Homework 5: Models and Model Comparisons

Max Campbell

Task 1: Conceptual Questions

- What is the purpose of using cross-validation when fitting a random forest model?

Generally, when fitting a random forest model, we want to randomly subset predictors to make sure that no predictor overwhelms the model and significantly effects the output. As such, we need to choose how many predictors are subset each time. This is where cross-validation comes into play. By implementing cross-validation and obtaining the most optimized tuning parameter we can improve our predictions.

- Describe the bagged tree algorithm.

In the bagged tree algorithm, we begin by bootstrapping some number of samples B. From there, we can fit a tree to each sample to have B number of trees. Then, from there, we can find response for each tree and then combine them to aggregate a final prediction. In practice, this combination is typically the average of all the sample predictions.

- What is meant by general linear model?

A general linear model is the family of models that we use more conventionally. For example, SLRs, MLRs, ANOVA and ANCOVA models all fall under this family (so long as all effects are fixed and not random).

- When fitting a multiple linear regression model, what does adding an interaction term do? That is, what does it allow the model to do differently as compared to when it is not included in the model?

The main reason to include an interaction variable is to allow the effect of one variable to be dependent on another variable. For example, in a plant growth study, the amount of sunlight and temperature may be heavily related to one another, so an interaction term may be beneficial to include in the model.

- Why do we split our data into a training and test set?

Splitting the model into a training set and test set allows us to see how well the model may perform if it was tasked with prediction future observations. By having test data, we can compare a model's predictions to the actual outcome to measure performance.

Task 2: Data Prep